Computationally-Efficient Sparsely-Connected Multi-Output Neural Networks for IM/DD System Equalization

Zhaopeng Xu*, Shuangyu Dong, Honglin Ji, Jonathan H. Manton, and William Shieh
Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia
**zhaopeng@student.unimelb.edu.au*

Abstract: Low-complexity sparsely-connected multi-output neural networks are proposed for equalization in a 50-Gb/s 25-km PAM4 IM/DD system. Compared with traditional fully-connected single-output counterparts, a gross complexity reduction of 60.4%/56.7% can be achieved with 2-layer FNN/C-FNN architecture. © 2022 The Author(s)

1. Introduction

The unprecedented popularity of data centers has led to an ever-increasing demand for short-reach optical interconnects. For short-reach applications, intensity-modulated directly-detected (IM/DD) systems are widely adopted due to the simple structure, low power consumption and low cost [1]. However, IM/DD systems suffer from nonlinear impairments owing to the mixture of chromatic dispersion and square-low direct detection [2]. The intrinsic nonlinear characteristics of low-cost lasers such as directly modulated lasers (DML) also significantly deteriorate the system performance [3,4]. A lot of digital signal processing (DSP) methods have been proposed to settle the nonlinear equalization issue, of which various neural network (NN)-based equalizers [5-7] attract the most attention since they can achieve better performance than traditional equalization schemes. However, NNs are usually computationally intensive, which go against the low power consumption trend of optical interconnects and hinder their practical implementation in real-time. Therefore, it is highly desirable to reduce the network complexity while upholding the system performance.

It has been proved that a proportion of weights inherent in NN-based nonlinear equalizers are redundant, and pruning can be adopted to cut off the insignificant weights [8,9]. In this paper, we propose multi-symbol equalization [10] combined with weight pruning to reduce NN complexity as much as possible. By increasing the number of outputs, the NN weights are shared for parallel symbol prediction, thereby reduce the number of multiplications required per received symbol [10]. Although multi-symbol equalization has already reduced computational complexity, we find there is still redundancy in specific weights, which make it possible for further complexity reduction. A DML based IM/DD experiment is demonstrated, which transmit 50-Gb/s pulse amplitude modulation (PAM)-4 signal over 25-km standard single mode fiber (SSMF) in C-band. Two different architectures of NNs, i.e., the simple feedforward NN (FNN) and cascade-FNN (C-FNN) [7], are selected for receiver-end equalization with the proposed multi-symbol prediction and pruning scheme. Compared with traditional fully-connected single-output FNN and C-FNN, the proposed method helps achieve a gross complexity reduction of 60.4% and 56.7% respectively, without degrading the original system bit-error-rate (BER) performance. The significant relief of complexity leads to only a few tens of multiplications per received symbol (57.0 for FNN and 68.8 for C-FNN), making NNs feasible for real-time implementation.

2. Sparsely-connected multi-output FNN/C-FNN receiver and experimental setup

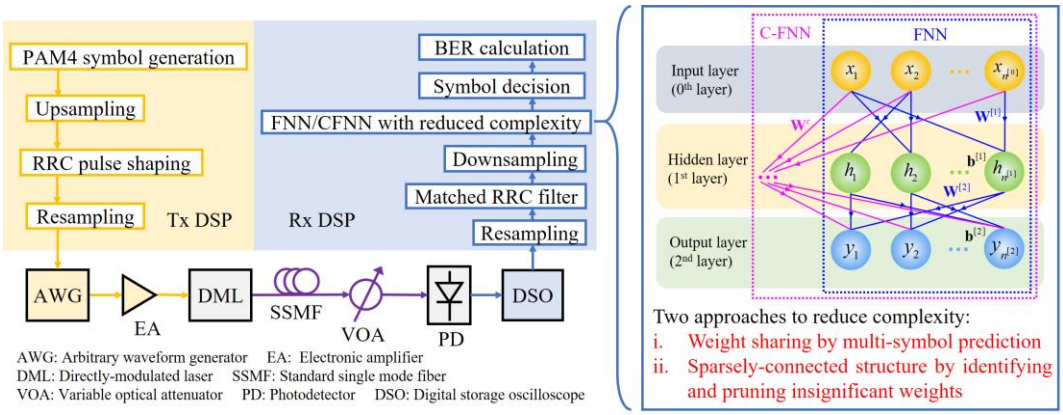

Fig. 1. Experimental setup of a 50-Gb/s 25-km PAM4 IM/DD link and the schematic of FNN/C-FNN with reduced complexity.

Fig. 1 illustrates the experimental setup of a DML-based IM/DD link which transmit 50-Gb/s PAM4 signal over 25-km SSMF. The schematic of Rx-end NN-based DSP with reduced complexity is also shown in the figure. At the transmitter side, the PAM4 signal is generated by a 92-GSa/s arbitrary waveform generator (AWG) with Tx DSP. The signal is then amplified by an electronic amplifier (EA) with 17-dB gain. A 16-GHz DML is used to convert the electrical signal into optical domain, and the optical signal is transmitted over 25-km SSMF. At the receiver side, a variable optical attenuator (VOA) is first employed to adjust the received optical power (ROP), and a 43-GHz photodetector (PD) is used to perform optical-electrical conversion. Finally, the received signal is captured by an 80-GSa/s digital storage oscilloscope (DSO), followed by Rx DSP.

Low-complexity FNN and C-FNN-based equalizers are proposed in Rx DSP to mitigate nonlinear system impairments. Compared with traditional NN-based equalizers, the computational complexity reduction approach is two-fold. First, the idea of multi-symbol prediction is adopted which produces multi-output NNs. In this way, part of the weights and biases can be shared when processing different symbols at the same time. Second, pruning technique gets involved to form a sparsely-connected NN structure. By identifying and cutting off insignificant weights, complexity can be further reduced. As shown in the figure, the number of NN input, hidden and output neurons is denoted by $n^{[0]}$, $n^{[1]}$, and $n^{[2]}$ respectively. For the i-th layer (i=1,2), the weight matrix $\mathbf{W}^{[i]}$ is a $n^{[i]} \times n^{[i-1]}$ matrix consists of all the weight coefficients connected from the (i-1)-th layer to the i-th layer and the bias matrix $\mathbf{b}^{[i]}$ is a $n^{[i]} \times 1$ vector includes all the biases of the i-th layer. The cascade structure of C-FNN is controlled by a $1 \times n^{[0]}$ weight vector $\mathbf{W}^c$. Assuming the $n^{[0]}$ NN inputs are denoted by $\mathbf{x}$, the $n^{[0]}$ NN outputs are denoted by $\mathbf{y}$, and the activation function of the i-th layer is denoted as $f^{[i]}(\cdot)$, the forward propagation step of NN, which is also the equalization process, can be expressed as

$$\mathbf{y} = f^{[2]}\left(\tilde{\mathbf{W}}^{[2]}\mathbf{h} + \mathbf{b}^{[2]}\right). \tag{1}$$

For FNN, we have $\mathbf{h} = f^{[1]}(\mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]})$, $\tilde{\mathbf{W}}^{[2]} = \mathbf{W}^{[2]}$, and for C-FNN, we have $\mathbf{h} = [(f^{[1]}(\mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}))^T, \mathbf{x}^T]^T$, $\tilde{\mathbf{W}}^{[2]} = [\mathbf{W}^{[2]}, \mathbf{W}^c]$. The complexity of FNN/C-FNN can be represented by the required number of multiplications per received symbol, which is denoted by N_{mul}. Considering the fully-connected NN structure, we have $N_{mul} = (n^{[0]}n^{[1]} + n^{[1]}n^{[2]})/n^{[2]}$ for FNN and $N_{mul} = (n^{[0]}n^{[1]} + n^{[1]}n^{[2]} + n^{[0]}n^{[2]})/n^{[2]}$ for C-FNN, which also equals the number of connections per output of each NN. After weight pruning, the entries in the weight matrices lower than a pre-defined weight threshold are set as 0, which means that the number of connections can be proportionally removed, thereby offering a more efficient computation. The inputs of both FNN and C-FNN contain the currently received $n^{[2]}$ symbols and the same number of past and post symbols, while the NN outputs correspond to the equalized $n^{[2]}$ symbols. We also note that to maintain BER and ensure a fair comparison, the input length of multi-output NNs should cover at least the same channel memory as that of single-output NNs for each output symbol. In other words, as $n^{[2]}$ increases, $n^{[0]}$ should also be increased by the same amount. For parameter settings of both NNs, tanh activation function is used in the hidden layer and a linear function is selected for the output layer. 20000 symbols are randomly selected to train the FNN/C-FNN, while another 1.2 million symbols are used as testing data.

3. Results and discussions

Before applying pruning, we first investigate on the effect of the number of NN outputs $n^{[2]}$ and determine a proper selection $n^{[2]}$ that has the lowest complexity. Fig. 2(a) depicts the N_{mul} threshold (lowest N_{mul} required to achieve the best BER performance) versus $n^{[2]}$ for FNN/C-FNN when the ROP is set at -1 dBm. Note that for both single-output FNN and C-FNN, at least 15 inputs and 9 hidden neurons are required for the best BER performance (around 3.8×10^{-3}), involving 144 and 159 multiplications respectively [10] which is shown by the black dashed line in Fig. 2(a). It can be observed that as $n^{[2]}$ increases, the minimum N_{mul} decreased at first, and then increases. For FNN,

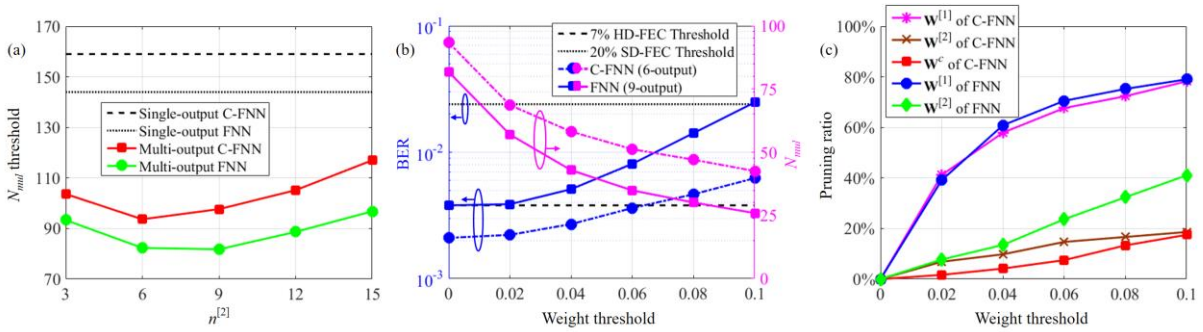

Fig. 2. (a) N_{mul} threshold versus $n^{[2]}$ for single/multi-output FNN/C-FNN; (b) BER and N_{mul} versus weight threshold for FNN (9-output) and C-FNN (6-output); (c) Pruning ratio of different types of weights in FNN (9-output) and C-FNN (6-output) versus weight threshold.

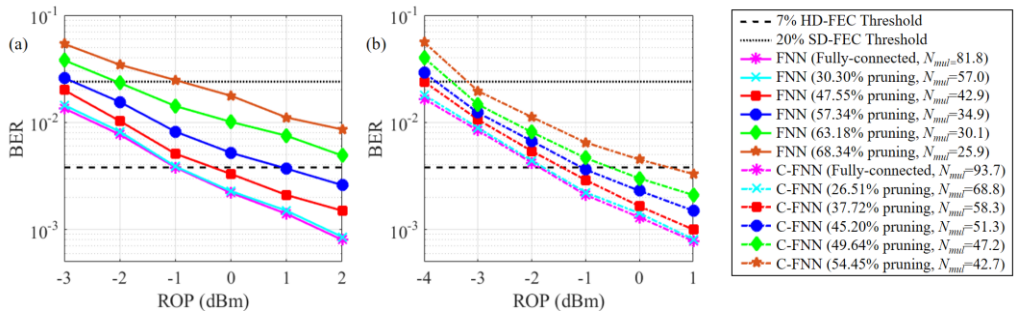

Fig. 3. BER versus ROP under different pruning ratio using (a) FNN (9-output) and (b) C-FNN (6-output).

the optimal $n^{[2]}$ is 9 ($n^{[1]}$=23, N_{mul}=81.8) while for C-FNN, the best $n^{[2]}$ is 6 ($n^{[1]}$=17, N_{mul}=93.7). Fig. 2(b) shows the BER and N_{mul} versus weight threshold for the best multi-output NNs (9-output FNN and 6-output C-FNN). As the weight threshold becomes larger, more weights are pruned, resulting in an increase of BER performance and a decrease of N_{mul} for both FNN and C-FNN. However, when pruning at a threshold of 0.02, the BER performance shows almost no difference compared with the fully-connected case, allowing a reduction of the number of connections without lowering BER. For the worst FNN case, the BER can still be lower than the 20% soft-decision forward error correction (SD-FEC) threshold even when the weight threshold is increased to 0.1, where only 25.9 multiplications is required. Fig. 2(c) present the pruning ratio of each weight matrix versus weight threshold. We can see that pruning mainly take place in the input-hidden layer weights $\mathbf{W}^{[1]}$, while for the more important hidden-output layer connections and cascade connections, less weights can be pruned.

The BER performance versus ROP under different pruning ratio of 9-output FNN and 6-output C-FNN are shown in Fig. 3(a) and 3(b) respectively. The pruning ratio or the complexity reduction ratio listed in the legend is calculated based on the weight thresholds selected in Fig. 2(b) and 2(c). Although $n^{[0]}$, $n^{[1]}$, and $n^{[2]}$ selected for FNN/C-FNN are the minimum values required for the best BER, experimental results show that there is still room for weight pruning. Compared to the fully-connected multi-output NNs, about 30.3% and 26.5% pruning reduction can be achieved for FNN and C-FNN respectively, while still upholding the BER performance. If we further relax the BER requirement to SD-FEC threshold, more than half of the NN connections can be removed. In a nutshell, multi-symbol equalization and pruning can significantly reduce the complexity without degrading BER. Only 57.0/68.8 multiplications per symbol is needed for FNN/C-FNN aided by the sparsely-connected multi-output structure, achieving a gross complexity reduction of 60.4%/56.7% compared with fully-connected single-output counterpart which requires 144/159 multiplications per symbol.

4. Conclusion

Multi-symbol equalization and weight pruning has been proposed and validated with a 2-layer FNN/C-FNN architecture which jointly reduce the computational complexity required for nonlinear equalization in IM/DD links. Compared with traditional fully-connected single-output FNN/C-FNN, an overall complexity reduction of 60.4%/56.7% is achieved in a 50-Gb/s 25-km PAM4 IM/DD system without degrading system BER performance.

5. References

[1] X. Pang, O. Ozolins, R. Lin, L. Zhang, A. Udalcovs, L. Xue, R. Schatz, U. Westergren, S. Xiao, W. Hu, G. Jacobsen, S. Popov, and J. Chen, "200 Gbps/lane IM/DD technologies for short reach optical interconnects," J. Lightw. Technol. **38**(2), 492-503 (2019).

[2] K. Zhong, X. Zhou, J. Huo, C. Yu, C. Lu, and A. P. K. Lau, "Digital signal processing for short-reach optical communications: A review of current technologies and future trends," J. Lightw. Technol. **36**(2), 377-400 (2018).

[3] D. Che, Y. Matsui, R. Schatz, G. Raybon, V. Bhatt, M. Kwakernaak, and T. Sudo, "Long-term reliable >200-Gb/s directly modulated lasers with 800GbE-compliant DSP," in Proc. OFC (2021), paper F3A.3.

[4] Y. Yu, M. R. Choi, T. Bo, Z. He, Y. Che, and H. Kim, "Low-complexity second-order Volterra equalizer for DML-based IM/DD transmission system," J. Lightw. Technol. **38**(7), 1735 - 1746 (2020).

[5] C. Häger and H. D. Pfister, "Nonlinear interference mitigation via deep neural networks," in Proc. OFC (2018), paper W3A.4.

[6] B. Karanov, M. Chagnon, F. Thouin, T. A. Eriksson, H. Bülow, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-end deep learning of optical fiber communications," J. Lightw. Technol. **36**(20), 4843-4855 (2018).

[7] Z. Xu, C. Sun, T. Ji, H. Ji, and W. Shieh, "Cascade recurrent neural network enabled 100-Gb/s PAM4 short-reach optical link based on DML," in Proc. OFC (2020), paper W2A.45.

[8] G. S. Yadav, C. Y. Chuang, K. M. Feng, J. Chen, and Y. K. Chen., "Computation efficient sparse DNN nonlinear equalization for IM/DD 112 Gbps PAM4 inter-data center optical interconnects," Opt. Lett. **46**(9), 1999-2002 (2021).

[9] L. Ge, W. Zhang, C. Liang, and Z. He, "Compressed neural network equalization based on iterative pruning algorithm for 112-Gbps VCSEL-enabled optical interconnects," J. Lightw. Technol. **38**(6), 1323-1329 (2020).

[10] Z. Xu, S. Dong, J. H. Manton, and W. Shieh, "Low-complexity multi-task learning aided neural networks for equalization in short-reach optical interconnects," J. Lightw. Technol. DOI: 10.1109/JLT.2021.3117687.